

Performance Analysis of Multi-Server based on Processor-Sharing Queue

Xiaofeng Zhang*, Baoqun Yin* and Hao Shi*

*Department of Automation University of Science and Technology of China, Hefei, China
zhxf325@mail.ustc.edu.cn, bqyin@ustc.edu.cn, haoshi@mail.ustc.edu.cn

Abstract—In the last few decades, the performance analysis of multi-server has become a research hotspot. We consider a distributed server system in which heterogeneous servers operating under the processor sharing(PS) discipline. In this system, when a stream of users' requests arrive at a dispatcher, it will dispatch each user to different servers respectively. In this paper, we analyze the performance of a single server using the finite capacity M/M/1/N-PS model. For this model, two exact expressions are derived for both user's average service time and average waiting time. Then we analyze the performance of the system by adopting different assignment strategies. Finally, considering the limitation of the service time and the waiting time, we give an assignment strategy in order to obtain the best performance.

Keywords—Network performance, Multi-Server system, assignment strategy, Processor-Sharing Queue, M/M/1/N-PS model

I. INTRODUCTION

With the development of the Internet, people are more like to rely on the web for their daily activities such as e-commerce, online banking, stock trading, reservation, and product merchandising. Consequently, it became evident that these popular sites must be offered with superior performance in order to retain existing users and attract new ones [1]. So performance analysis of servers is quite necessary.

Some work has been done to test server performance. Menasce presented a good review of server performance testing in [2]. There exist several commercial and open-source testing tools. In [3] the author presented an overview of some of these tools. Simulation models also exist for server performance testing [4]. However these work lacks of theory analysis.

Some work has focused on performance modeling and analysis of servers. Slothouber [5] modeled the server as an open queueing network. Dilley et al. [6, 7] used layered queueing models regarding the performance evaluation of a server in the Internet and Intranet. Reeser and Hariharan [8] presented an end-to-end analytical queueing model for dynamic server-side computing following a Poisson arrival process. In [9] the author proposed a Markovian queueing model with a variable number of servers to analyze the performance of an Apache server. Aljohani [10] introduced a simple queueing model to analyze the performance metrics of server under varying traffic loads. Esa Hyytiä [11] considered dispatching problem

in a distributed server system. The author proposed some assignment strategies and compared them with simulation.

Although much work about analyzing and testing servers performance has been done, rare work derives the expressions of the average service time, the average waiting time and the maximum numbers of concurrent users under PS discipline. However these are important performance targets. In this paper, we consider the system illustrated in Fig. 1. The servers are operated under PS discipline. When a user arrives, it will be dispatched to one of the servers. The user will be served immediately or later according to the number of users being served is more or less than the maximum number of concurrent users.

The rest of this paper is organized as follows. In section II, we analyze the performance of a single server based on an M/M/1/N-PS model. In section III, performance of multi-server system is analyzed under different assignment strategies. Section IV contains the numerical examples. Finally section V concludes the paper.

II. PERFORMANCE ANALYSIS OF A SINGLE SERVER

In this section we analyze an M/M/1/N-PS model and derive the expressions of the average service time and the average waiting time. The result is utilized later in section III to analyze the performance of multi-server.

Let λ denote the Poissonian arrival rate and $1/\mu$ denote the mean of the exponentially distributed service time. Furthermore, $\rho = \lambda/\mu$ is the offered load assumed to be less than

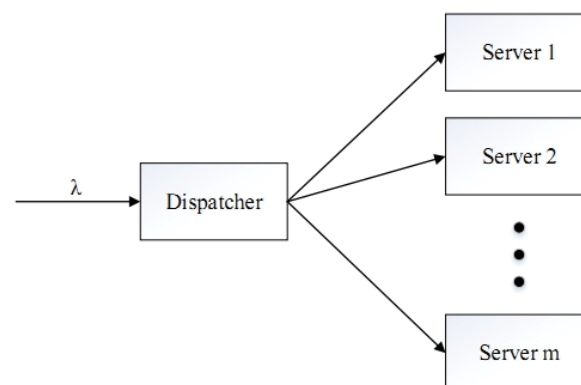


Fig. 1. The System Architecture.

one, $\rho < 1$. We assume that one server can serve at most N users at the same time. In the other words, the maximum number of concurrent users is N . When a user arrives at the system, it will be served immediately if there are less than N users being served. Otherwise, it will wait. We assume the queue's capacity is infinite. This is an M/M/1/N-PS model.

Let J indicate service time the user need. And when the system is in the steady state, K denotes the number of users in the system, W denotes the user's waiting time and S denotes the user's service time.

Because the departure process is same as M/M/1 queue model, the stable distribution of the number of users is

$$P\{K = n\} = (1 - \rho)\rho^n, \quad n = 0, 1, 2, \dots \quad (1)$$

A. Analysis of the average service time

Assume the service time a user need is x , when it arrives at the system there are n users being served. Let $F_n(x, t)$ denote conditional distribution function of the service time S under $K = n, J = x$ when the system is in the steady state. So,

$$F_n(x, t) = P\{S \leq t | K = n, J = x\}, \\ n = 0, 1, 2, \dots, x > 0.$$

Let $f_n(x, t)$ denote conditional probability density function,

$$f_n(x, t)dt = F_n(x, t + dt) - F_n(x, t), \\ n = 0, 1, 2, \dots, x > 0.$$

Next, we analyze $f_n(x, t)$ under two conditions based on a arriving user need to wait or not.

1) $n = 0, 1, \dots, N-1$: When a user arrives to the system, it will be served immediately in this condition. Base on the system model, the conditional probability density function is same as finite capacity M/M/1-PS model. Base on [12], we can get

$$\begin{aligned} \frac{\partial}{\partial t} f_n(x, t) &= -\frac{1}{n+1} \frac{\partial}{\partial x} f_n(x, t) + \lambda [f_{n+1}(x, t) - f_n(x, t)] \\ &\quad + \frac{\mu n}{n+1} [f_{n-1}(x, t) - f_n(x, t)], \\ n &= 0, 1, 2, \dots, N-2, x > 0. \end{aligned} \quad (2)$$

Its boundary condition is

$$\begin{aligned} \frac{\partial}{\partial t} f_{N-1}(x, t) &= \frac{\mu(N-1)}{N} [f_{N-2}(x, t) - f_{N-1}(x, t)] \\ &\quad - \frac{1}{N} \frac{\partial}{\partial x} f_{N-1}(x, t). \end{aligned} \quad (3)$$

The conditional mathematical expectation of S under $K = n, J = x$ is

$$\begin{aligned} e_n(x) &\triangleq E\{S | K = n, J = x\} \\ &= \int_0^\infty t f_n(x, t) dt, \quad n = 0, 1, 2, \dots, N-1. \end{aligned} \quad (4)$$

Based on (2), (3) and (4), we can get

$$\begin{aligned} -\frac{1}{n+1} \frac{d}{dx} e_n(x) + \lambda [e_{n+1}(x) - e_n(x)] + \frac{\mu n}{n+1} \times \\ [e_{n-1}(x) - e_n(x)] = -1, \quad n = 0, 1, 2, \dots, N-2. \end{aligned} \quad (5)$$

Its boundary condition is

$$\begin{cases} -\frac{1}{N} \frac{d}{dx} e_{N-1}(x) + \frac{\mu(N-1)}{N} \times [e_{N-2}(x) - e_{N-1}(x)] = -1, \\ e_n(0) = 0, n = 0, 1, \dots, N-1. \end{cases} \quad (6)$$

2) $n = N, N+1, \dots$: In this condition a user must wait when it arrives at the system. According to the above,

$$F_n(x, t) = F_{N-1}(x, t), \quad n = N, N+1, \dots \quad (7)$$

The conditional mathematical expectation of S under $K = n, J = x$ satisfies

$$\begin{aligned} E\{S | K = n, J = x\} &= E\{S | K = N-1, J = x\}, \\ n &= N, N+1, \dots \end{aligned} \quad (8)$$

Introducing the generating function

$$\begin{aligned} T(x, z) &= \sum_{n=0}^{N-1} e_n(x) z^n \\ &\quad + \sum_{n=N}^{\infty} E\{S | K = N-1, J = x\} z^n. \end{aligned} \quad (9)$$

According to (1), (9) and [12], we get the conditional mathematical expectation of S under $J = x$ is

$$E\{S | J = x\} = (1 - \rho) T(x, \rho).$$

Use the Laplace transform used in [13], according to (2)-(8), we can get

$$\begin{aligned} E\{S | J = x\} &= \frac{x}{1 - \rho} [1 - \rho^N + N\rho^N(\rho - 1)] + N\rho^N x \\ &= \frac{1 - \rho^N}{1 - \rho} x. \end{aligned}$$

Because the service time a user need follow exponential distribution with parameter μ ,

$$P\{J \leq x\} = 1 - e^{-\mu x}, \quad x > 0.$$

Then we can get the average service time is

$$E(S) = E[E(S | J)] = \frac{1 - \rho^N}{1 - \rho} E(J) = \frac{1 - \rho^N}{\mu(1 - \rho)}. \quad (10)$$

B. Analysis of the average waiting time

As the same with analysis of the mean service time, the average waiting time is analyzed under two conditions. If there are less than N users in the system, a user will not need to wait when it arrives at the system. So,

$$\begin{aligned} P\{W = 0 | K = n, J = x\} &= 1, \\ n &= 0, 1, \dots, N-1, x > 0. \end{aligned} \quad (11)$$

When $n = N, N+1, \dots$, the conditional mathematical expectation of W under $K = n, J = x$ is

$$E(W | K = n, J = x) = (n - N + 1)E(S | J = x). \quad (12)$$

As the same with the service time, according to (1), (11) and (12),

$$\begin{aligned} E(W|J=x) &= \sum_{n=N}^{\infty} (1-\rho) \rho^n (n-N+1) E(S|J=x) \\ &= \frac{\rho^N}{1-\rho} E(S|J=x). \end{aligned}$$

So, the average waiting time is

$$\begin{aligned} E(W) &= E[E(W|J)] \\ &= \frac{\rho^N}{1-\rho} E(S) = \frac{\rho^N (1-\rho^N)}{\mu(1-\rho)^2}. \end{aligned} \quad (13)$$

III. PERFORMANCE ANALYSIS OF MULTI-SERVER

In this section, the results in previous section are utilized to solve multi-server problem. We derive the expressions of the average service time and the average waiting time under different assignment strategies. Then we give the maximum number of concurrent users under the limitation of the average service time and the average waiting time. Some numerical examples are given in the next section.

Assume that the number of servers in the system is m , and the service rate of server i is μ_i . The maximum number of concurrent users of server i is N_i . Let λ denote the Poissonian arrival rate. A user is dispatched to one of the m servers independently in random using a probability distribution p_i , $i = 1, \dots, m$. So for each server, it is an M/M/1/N-PS model mentioned in previous section.

The arrival rate of server i is

$$\lambda_i = p_i \lambda.$$

And the offered load of server i is

$$\rho_i = \frac{\lambda_i}{\mu_i} = \frac{p_i \lambda}{\mu_i}.$$

In this paper, we consider two random assignment strategies: EQU- ρ and EQU-opt.

A. EQU- ρ

One rational choice for the dispatching policy is to balance the load equally among the servers, so that $\rho_i = \rho_j$ for all i, j . To this end, EQU- ρ uses probability distribution

$$p_i = \frac{\mu_i}{\sum_i \mu_i},$$

which then gives

$$\rho_i = \frac{\lambda_i}{\mu_i} = \frac{\lambda}{\sum_i \mu_i} = \rho. \quad (14)$$

The resulting system is stable as long as the offered load ρ is less than one.

Then according to (10) and (14), the average service time of server i is

$$E_i(S) = \frac{1-\rho_i^{N_i}}{\mu_i(1-\rho_i)}, \quad i = 1, 2, \dots, m.$$

According to (13) and (14), the average waiting time of server i is

$$E_i(W) = \frac{\rho_i^{N_i} (1-\rho_i^{N_i})}{\mu_i(1-\rho_i)^2}, \quad i = 1, 2, \dots, m.$$

So the average service time of the whole system is

$$E(S) = \sum_{i=1}^m p_i E_i(S) = \frac{1}{\sum_{i=1}^m \mu_i} \sum_{i=1}^m \frac{1-\rho_i^{N_i}}{1-\rho_i}.$$

And the average waiting time of all users is

$$E(W) = \sum_{i=1}^m p_i E_i(W) = \frac{1}{\sum_{i=1}^m \mu_i} \sum_{i=1}^m \frac{\rho_i^{N_i} (1-\rho_i^{N_i})}{(1-\rho_i)^2}.$$

B. EQU-opt

EQU-opt policy uses the probability distribution that minimizes the mean sojourn time which is sum of the average service time and the average waiting time. Based on before, we can get the mean sojourn time is

$$\begin{aligned} E(T) &= \sum_{i=1}^m p_i E_i(T) = \sum_{i=1}^m p_i (E_i(S) + E_i(W)) \\ &= \sum_{i=1}^m p_i \cdot \frac{(1-\rho_i^{N_i})(1-\rho_i+\rho_i^{N_i})}{\mu_i(1-\rho_i)^2}. \end{aligned} \quad (15)$$

Since every $E_i(T)$ is independent to other, we can calculate the value of N_i when $E_i(T)$ get the minimum. Let $a_N = (1-\rho^N)(1-\rho+\rho^N)$. So

$$a_{N+1} - a_N = \rho^{N+1}(1-\rho)(\rho^N + \rho^{N-1} - 1).$$

When $N = 1$, $a_2 - a_1 = \rho^3(1-\rho) > 0$, and when N is big enough, $a_{N+1} - a_N < 0$. So with the increasing of N , a_N first increase and then decrease. Obviously, $E_i(T)$ is same with a_N . Thus the $E_i(T)$ get the minimum when $N_i = 1$ or $N_i \rightarrow \infty$. However $\lim_{N_i \rightarrow \infty} E_i(T) = \frac{1}{\mu_i(1-\rho_i)}$ is equal to the value of $E_i(T)$ when $N_i = 1$. So (15) can be simplified to

$$E(T) = \sum_{i=1}^m p_i E_i(T) = \sum_{i=1}^m \frac{p_i}{\mu_i(1-\rho_i)} = \sum_{i=1}^m \frac{p_i}{\mu_i - p_i \lambda}.$$

The optimal probability distribution can be obtained with aid of Lagrange method. We introduce the Lagrange multiplier t . Let

$$F(p_i, t) = \sum_{i=1}^m \frac{p_i}{\mu_i - p_i \lambda} - t \left(\sum_{i=1}^m p_i - 1 \right).$$

According to Lagrange method, we can get

$$\begin{cases} \frac{\partial F}{\partial p_i} = \frac{\mu_i}{(\mu_i - p_i \lambda)^2} - t = 0 \\ \frac{\partial F}{\partial t} = \sum_{i=1}^m p_i - 1 = 0. \end{cases} \quad (16)$$

Solving (16), we can get

$$p_i = \frac{1}{\lambda} \left(\mu_i - \frac{\sqrt{\mu_i}}{\beta} \right),$$

where $\beta = \frac{\sum_{i=1}^m \sqrt{\mu_i}}{(\sum_{i=1}^m \mu_i) - \lambda}$.

So the offered load for server i is

$$\rho_i = \frac{\lambda_i}{\mu_i} = 1 - \frac{1}{\beta \sqrt{\mu_i}}. \quad (17)$$

According to (10) and (17), the average service time of server i is

$$E_i(S) = \frac{1 - \rho_i^{N_i}}{\mu_i(1 - \rho_i)}, \quad i = 1, 2, \dots, m.$$

According to (13) and (17), the average waiting time of server i is

$$E_i(W) = \frac{\rho_i^{N_i}(1 - \rho_i^{N_i})}{\mu_i(1 - \rho_i)^2}, \quad i = 1, 2, \dots, m.$$

Then the average service time of the whole system is

$$E(S) = \sum_{i=1}^m p_i E_i(S) = \frac{1}{\lambda} \left(\mu_i - \frac{\sqrt{\mu_i}}{\beta} \right) \sum_{i=1}^m \frac{1 - \rho_i^{N_i}}{\mu_i(1 - \rho_i)}.$$

And the average waiting time of all users is

$$\begin{aligned} E(W) &= \sum_{i=1}^m p_i E_i(W) \\ &= \frac{1}{\lambda} \left(\mu_i - \frac{\sqrt{\mu_i}}{\beta} \right) \sum_{i=1}^m \frac{\rho_i^{N_i}(1 - \rho_i^{N_i})}{\mu_i(1 - \rho_i)^2}. \end{aligned}$$

C. Performance optimization

In the real system, the average service time and the average waiting time are limited. In this part, the optimal probability distribution is given to get the best performance with the limitation of the average service time and the average waiting time. We take the mean sojourn time and the maximum number of concurrent users as the performance target. Assume that the average service time and the average waiting time of server i must be less than CS_i and CW_i . Then we get a multi-objective optimization problem,

$$\begin{aligned} \min_{p_i} E(T), \max_{N_i} N &= \sum_{i=1}^m N_i \\ s.t. \quad &\begin{cases} E_i(S) \leq CS_i \\ E_i(W) \leq CW_i \\ \sum_{i=1}^m p_i = 1 \\ 0 < p_i < 1 \end{cases}, \quad i = 1, 2, \dots, m. \end{aligned}$$

The NSGA-II mentioned in [14] can be used to get the probability distribution.

IV. NUMERICAL EXAMPLES

In this section, we give numerical results of a single server system to confirm the correctness of our analytical results. Then the numerical results of a two-server system are given to compare the performance of system under two different assignment strategies mentioned before. Finally, with the limitation of the average service time and the average waiting time, we get the probability distribution and confirm its correctness.

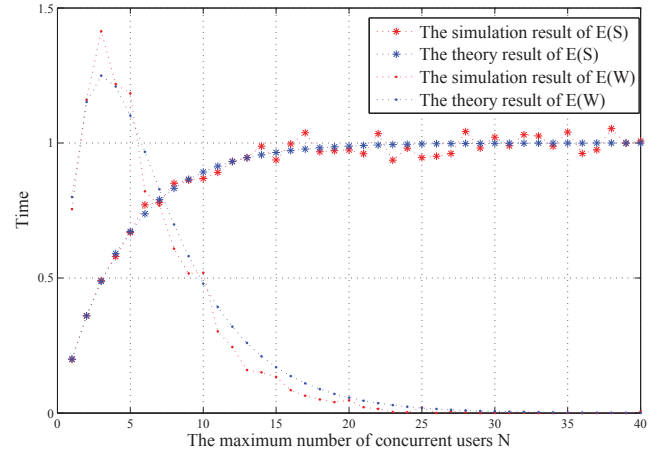


Fig. 2. The performance of a single-server system.

For the single server system, we set $\lambda = 4, \mu = 5$. Fig. 2 shows that how the average service time and the average waiting time change with increasing of the maximum number of concurrent users. Form the figure, we can see that the simulation results and the theory results are similar. With the increasing of the maximum number of concurrent users, the mean service time increases and reaches a settled value when the number of concurrent users tends to infinite. And the average waiting time decreases to 0 when the maximum number of concurrent users tends to infinite. This is an obvious result since none user need to wait when the maximum number of concurrent users tends to infinite.

We set $\lambda = 10, \mu_1 = 6, \mu_2 = 8$ for the two-server system. Fig. 3-6 show the performance of two servers under two different assignment strategies. The probability distribution is $p_1 = 0.43, p_2 = 0.57$ with EQU- ρ and $p_1 = 0.41, p_2 = 0.59$ with EQU-opt. From these four figures, we can see the average

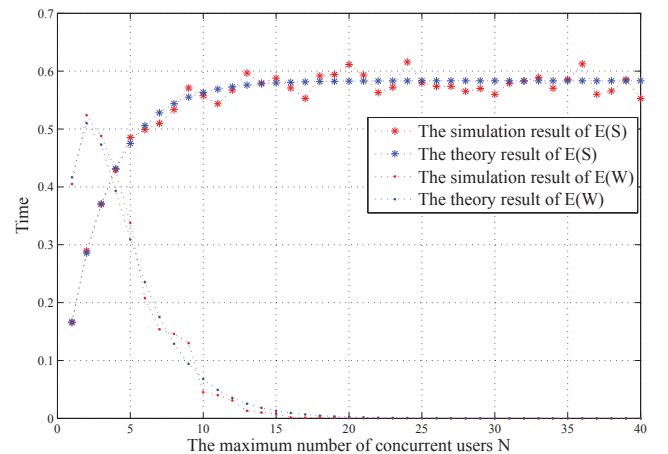


Fig. 3. The performance of server 1 under EQU- ρ .

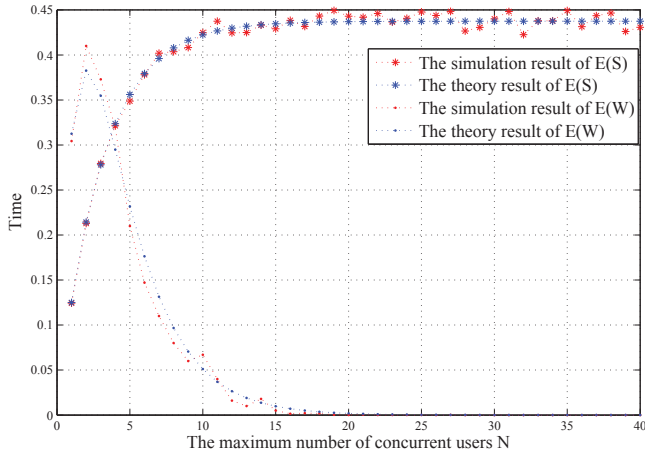


Fig. 4. The performance of server 2 under EQU- ρ .

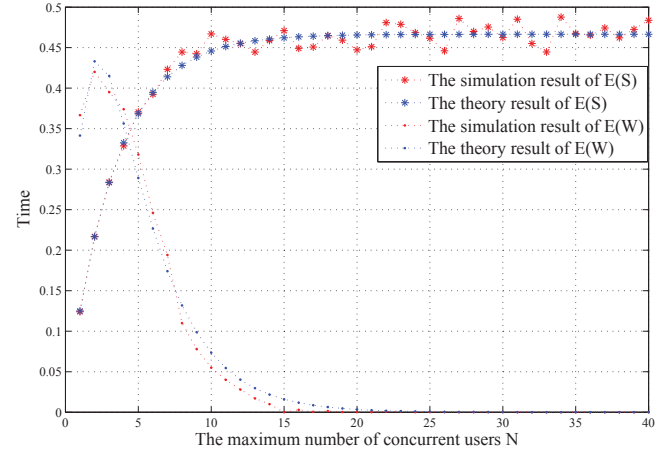


Fig. 6. The performance of server 2 under EQU-opt.

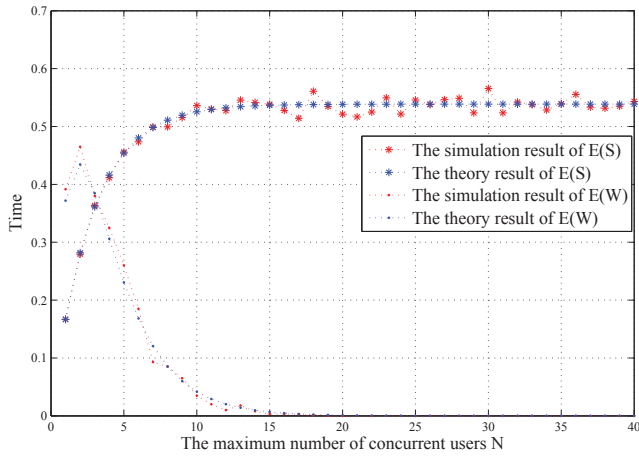


Fig. 5. The performance of server 1 under EQU-opt.

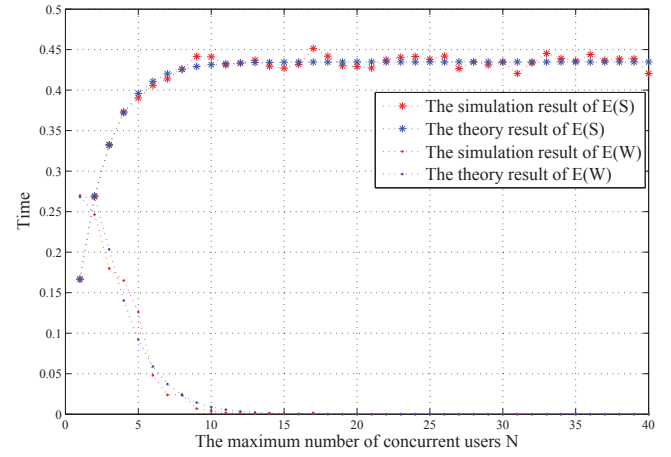


Fig. 7. The performance of server 1 under time limitation

service time for server1 is 0.59 and 0.53 with two different assignment strategies, and the average service time for server2 is 0.44 and 0.46 with two different assignment strategies when the maximum number of concurrent users is infinite. So the mean sojourn time is 0.5045 for EQU- ρ and 0.4887 for EQU-opt. Thus the performance of EQU-opt is better than EQU- ρ in the aspect of the mean sojourn time.

For the limitation of the average service time and the average waiting time, we set $\lambda = 10, \mu_1 = 6, \mu_2 = 8, ES_1 = 0.4, ES_2 = 0.3, EW_1 = 0.6, EW_2 = 0.5$. Using the NSGA-II method, we solve the multi-subject optimization problem and get $p_1 = 0.37, p_2 = 0.63, N_1 = 5, N_2 = 1$. From Fig. 7 and Fig. 8, we can see the average service time and the average waiting time correspond to the limitation, but the mean sojourn time is larger than EQU-opt and EQU- ρ .

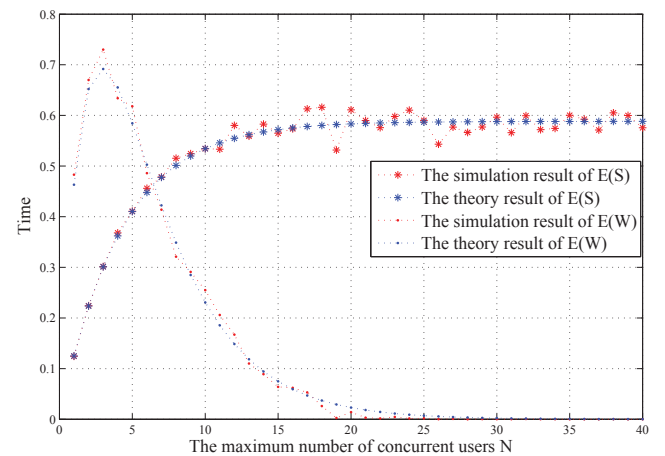


Fig. 8. The performance of server 2 under time limitation

V. CONCLUSIONS

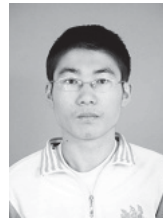
In this paper, we have studied the performance of multi-server under the PS discipline. In such system, arriving users are dispatched to one of the available servers, which then processes the given users in parallel according to the PS discipline. In order to analyze the performance of different assignment strategies, we study the performance of a single server based on M/M/1/N-PS model. For this model, we derived the expressions of the average service time and the average waiting time. Then we apply it to analyze the performance of two different assignment strategies and compare their performance by numerical examples. And find that the performance of EQU-opt is better than EQU- ρ in aspect of the mean sojourn time. Considering the limitations of the service time and the waiting time in the real system, we get a multi-object optimization problem with the target of maximize the number of concurrent users in whole system and minimize the mean sojourn time. The NSGA-II method is used to solve the problem. In numerical examples, the result is correspond to the limitation and the performance of this condition is worse than EQU-opt. The future work includes considering cooperation of different servers.

ACKNOWLEDGMENTS

This work is supported in part by the National Natural Science Foundation of China under grant Nos.61174124,61233003, in part by Research Fund for the Doctoral Program of Higher Education of China under grant No.20123402110029 and in part by Natural Science Research Program of the Anhui High Education Bureau of China under grant No.KJ2012A286.

REFERENCES

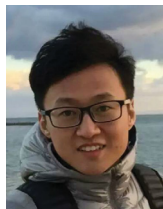
- [1] R. D. van der Mei, R. Hariharan, and P. Reeser, "Web server performance modeling", *Telecommunication Systems*, vol. 16, no. 3-4, pp. 361-378, 2001.
- [2] D. Menasce, "Load testing of web sites", *Internet Computing, IEEE*, vol. 6, no. 4, pp. 70-74, 2002.
- [3] J. Križanić, A. Grgurić, M. Mošmondor and P. Lazarevski, "Load testing and performance monitoring tools in use with AJAX based web applications", in *33rd International Convention on Information and Communication Technology, Electronics and Microelectronics*, Opatija, pp. 428-434, 2010.
- [4] L. Wells, S. Christensen, L. Kristensen, and K. Mortensen, "Simulation based performance analysis of web servers", in *ninth International Workshop on Petri Nets and Performance Models*, Aachen, pp. 59-68, 2001.
- [5] L. Slothouber, "A model of web server performance", in *fifth International World Wide Web Conference*, Paris, 1996.
- [6] J. Dilley, R. Friedrich, T. Jin and J. Rolia, "Web server performance measurement and modeling techniques", *Performance Evaluation*, vol. 33, no. 1, pp. 5-26, 1998.
- [7] J. Dilley, R. Friedrich, T. Jin and J. Rolia, "Measurement tools and modeling techniques for evaluating Web server performance", *Computer Performance Evaluation: Modelling Techniques and Tools*, vol. 1245, pp. 155-168, 1997.
- [8] P. Reeser, R. Hariharan, "An analytic model of web servers in distributed computing environments", *Telecommunications Systems*, vol. 21, no. 2-4, pp. 283-299, 2002.
- [9] T.V. Do, U.R. Krieger and R. Chakka, "Performance modeling of an Apache Web server with a dynamic pool of service processes", *Telecommun Systems*, vol. 39, no. 2, pp. 117-129, 2008.
- [10] A.M.D Aljohani, D.R.W Holton, and I. Awan, "Modeling and performance analysis of Scalable Web Servers Deployed on the Cloud", in *eighth International Conference on Broadband, Wireless Computing, Communication and Applications*, Compiegne, pp. 238-242, 2013.
- [11] Esa Hyytiä, Jorma Virtamo, Samuli Aalto and Aleksi Penttinen, "M/M/1-PS queue and size-aware task assignment", *Performance Evaluation*, vol. 68, no. 11, pp. 1136-1148, 2011.
- [12] Knessl C, "On finite capacity processor-shared queues", *SIAM Journal on Applied Mathematics*, vol. 50, no. 1, pp. 264-287, 1990.
- [13] Knessl C, "On the sojourn time distribution in a finite capacity processor shared queue", *Journal of the ACM*, vol. 40, no. 5, pp. 1238-1301, 1993.
- [14] Abdullah Konak, David W. Coit and Alice E. Smith, "Multi-objective optimization using genetic algorithms: A tutorial", *Reliability Engineering and System Safety*, vol. 91, no. 9, pp. 992-1007, 2006.



Xiaofeng Zhang received the B.S. degree in automation from the University of Science and Technology of China, Hefei, China, in 2011. He is currently pursuing the Ph.D. degree in the Department of Automation, University of Science and Technology of China. His research interests include modeling, optimization and performance analysis of the networking.



Baoqun Yin received the B.S. degree in fundamental mathematics from Sichuan University, Chengdu, China, in 1985, his M.S. degree in applied mathematics from University of Science and Technology of China, Hefei, China, in 1993, and his Ph.D. degree in pattern recognition and intelligent system from University of Science and Technology of China, Hefei, China, in 1998. He is currently a Professor in the Department of Automation, University of Science and Technology of China, Hefei, China. His current research interests include discrete event dynamic systems, Markov decision processes, queuing systems, and network resource management and optimization.



Hao Shi received the B.S. degree in automation from the University of Science and Technology of China, Hefei, China, in 2012. He is currently pursuing the Ph.D. degree in the Department of Automation, University of Science and Technology of China. His research interests include manifold learning, dimension reduction.